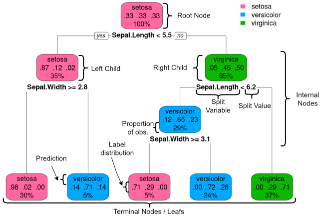
# Introduction to Machine Learning

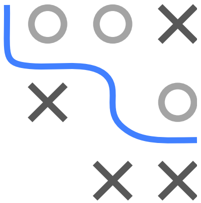
# Advanced Risk Minimization
# Loss functions and tree splitting




## Learning goals

- Tree splitting loss vs impurity:
- Bernoulli loss $\sim$ entropy splitting
- Brier score $\sim$ gini splitting

# RISK MINIMIZATION AND IMPURITY



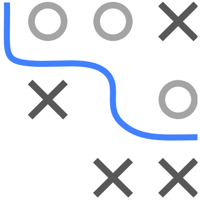
- Tree fitting: Find best way to split parent node $\mathcal{N}_0$ into child nodes $\mathcal{N}_1$ and $\mathcal{N}_2$, such that $\mathcal{N}_1 \cup \mathcal{N}_2 = \mathcal{N}_0$ and $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$
- Two options for evaluating how good a split is: Per node $\mathcal{N}$ compute the following:
  1. Compute impurity $\text{Imp}(\mathcal{N})$ directly from observations in $\mathcal{N}$
  2. Fit optimal constant using loss function, sum up losses for $\mathcal{N}$
- Summarize on split level:
  1. Weighted average ($n_0 = n_1 + n_2$ are number of obs in nodes)

$$\text{Imp(split)} = \frac{n_1}{n_0}\text{Imp}(\mathcal{N}_1) + \frac{n_2}{n_0}\text{Imp}(\mathcal{N}_2)$$

  2. Sum of individual losses

$$\mathcal{R}(\text{split}) = \mathcal{R}(\mathcal{N}_1) + \mathcal{R}(\mathcal{N}_2)$$

# BERNOULLI LOSS MIN = ENTROPY SPLITTING



**Claim:** Using entropy in (1) is equivalent to using Bernoulli loss in (2)
**Proof:** $\mathcal{N} \subseteq \mathcal{D}$ denotes subset of observations in that node.
Risk $\mathcal{R}(\mathcal{N})$ of node $\mathcal{N}$ w.r.t. (multiclass) Bernoulli loss

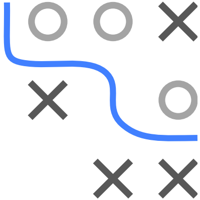
$$L\left(y, \pi(\mathbf{x})\right) = -\sum_{k=1}^{g} [y = k] \log\left(\pi_k(\mathbf{x})\right)$$

$\Rightarrow$ Optimal constant per node $\pi_k^{(\mathcal{N})} = \frac{1}{n_{\mathcal{N}}} \sum\limits_{(\mathbf{x},y)\in\mathcal{N}} [y = k]$.

Entropy of node $\mathcal{N}$:

$$\text{Imp}(\mathcal{N}) = -\textstyle\sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})}$$

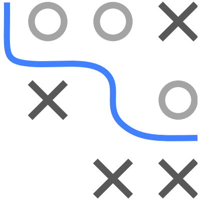
# RISK MINIMIZATION AND IMPURITY



$$
\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{(\mathbf{x},y)\in\mathcal{N}} \left( -\sum_{k=1}^{g} [y=k] \log \pi_k(\mathbf{x}) \right) \\
&= -\sum_{k=1}^{g} \sum_{(\mathbf{x},y)\in\mathcal{N}} [y=k] \log \pi_k^{(\mathcal{N})} \\
&= -\sum_{k=1}^{g} \log \pi_k^{(\mathcal{N})} \\
&= -n_{\mathcal{N}} \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})} = n_{\mathcal{N}} \mathrm{Imp}(\mathcal{N}) \\
\Rightarrow \mathcal{R}(\text{split}) &= \mathcal{R}(\mathcal{N}_1) + \mathcal{R}(\mathcal{N}_2) = n_1 \mathrm{Imp}(\mathcal{N}_1) + n_2 \mathrm{Imp}(\mathcal{N}_2) \\
&= n_0 \left( \frac{n_1}{n_0} \mathrm{Imp}(\mathcal{N}_1) + \frac{n_2}{n_0} \mathrm{Imp}(\mathcal{N}_2) \right) = n_0 \mathrm{Imp}(\text{split})
\end{aligned}
$$

Bernoulli-risk of the split $\mathcal{R}(\text{split})$ is proportional to its entropy-impurity Imp(split), i.e.,
$\arg\min_{\text{split}} \mathcal{R}(\text{split}) = \arg\min_{\text{split}} \mathrm{Imp}(\text{split})$

# BRIER SCORE MINIMIZATION = GINI SPLITTING



**Claim:** Using Gini in (1) is equivalent to using Brier score in (2)
**Proof:**
Risk $\mathcal{R}(\mathcal{N})$ of node $\mathcal{N}$ w.r.t. (multiclass) Brier score

$$L(y, \pi(\mathbf{x})) = \sum_{k=1}^{g} ([y = k] - \pi_k(\mathbf{x}))^2$$

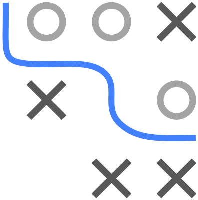
$\Rightarrow$ Optimal constant per node: $\pi_k^{(\mathcal{N})} = \frac{1}{n_{\mathcal{N}}} \sum\limits_{(\mathbf{x},y)\in\mathcal{N}} [y = k] = \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}$

($n_{\mathcal{N},k}$ is the number of class $k$ observations in node $\mathcal{N}$)

Gini index of node $\mathcal{N}$:

$$\text{Imp}(\mathcal{N}) = \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \left(1 - \pi_k^{(\mathcal{N})}\right)$$

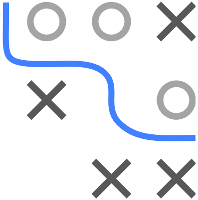
# BRIER SCORE MINIMIZATION = GINI SPLITTING



$$
\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{(\mathbf{x},y)\in\mathcal{N}} \sum_{k=1}^{g} \left([y=k] - \pi_k^{(\mathcal{N})}\right)^2 = \sum_{k=1}^{g} \sum_{(\mathbf{x},y)\in\mathcal{N}} \left([y=k] - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 \\
&= \sum_{k=1}^{g} \left( \sum_{(\mathbf{x},y)\in\mathcal{N}:\ y=k} \left(1 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 + \sum_{(\mathbf{x},y)\in\mathcal{N}:\ y\neq k} \left(0 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 \right) \\
&= \sum_{k=1}^{g} n_{\mathcal{N},k} \left(1 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2,
\end{aligned}
$$

since for $n_{\mathcal{N},k}$ observations the condition $y = k$ is met, and for the remaining $(n_{\mathcal{N}} - n_{\mathcal{N},k})$ observations it is not.

# BRIER SCORE MINIMIZATION = GINI SPLITTING



We further simplify the expression to

$$
\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{k=1}^{g} n_{\mathcal{N},k} \left( \frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left( \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 \\
&= \sum_{k=1}^{g} \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} (n_{\mathcal{N}} - n_{\mathcal{N},k} + n_{\mathcal{N},k}) \\
&= n_{\mathcal{N}} \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \cdot \left( 1 - \pi_k^{(\mathcal{N})} \right) = n_{\mathcal{N}} \text{Imp}(\mathcal{N}) \\
\Rightarrow \mathcal{R}(\text{split}) &= \mathcal{R}(\mathcal{N}_1) + \mathcal{R}(\mathcal{N}_2) = n_1 \text{Imp}(\mathcal{N}_1) + n_2 \text{Imp}(\mathcal{N}_2) \\
&= n_0 \left( \frac{n_1}{n_0} \text{Imp}(\mathcal{N}_1) + \frac{n_2}{n_0} \text{Imp}(\mathcal{N}_2) \right) = n_0 \text{Imp}(\text{split})
\end{aligned}
$$

Brier-risk of the split $\mathcal{R}(\text{split})$ is proportional to its gini-impurity Imp(split), i.e.,
$\arg\min_{\text{split}} \mathcal{R}(\text{split}) = \arg\min_{\text{split}} \text{Imp}(\text{split})$